CS 7545: Machine Learning Theory **Fall 2021**

Lecture 18

Name: Tyler LaBonte **Professor:** Santosh S. Vempala

1 Statistical Query Learning

We've covered the PAC model for learning, but it does not handle every possible scenario. In particular, we may be interested in the case when we have random errors in the labels. For example, $\ell(x) = f(x)$ with probability $1-\eta$ and $1-f(x)$ with probability η. Can we still learn the underlying concept in this case?

Suppose we are learning a logical OR of boolean variables. Let $p_i = \Pr[f(x) = 0 and x_i = 1]$. Then, we can let the output hypothesis h be all variables for which $p_i = 0$ and no variables for which $p_i > \frac{\epsilon}{n}$. So, we need to estimate each p_i to within additive error $\frac{\epsilon}{2n}$ to take a union bound. We have

$$p_i = \Pr[f(x) = 0 | x_i = 1] \cdot \Pr[x_i = 1]. \tag{1}$$

The right side is independent of the label noise. Call the left side q_i, then

$$\Pr_{\eta}[\ell(x) = 0 | x_i = 1] = (1-\eta)q_i + \eta(1-q_i) \tag{2}$$

$$= \eta + q_i(1-2\eta). \tag{3}$$

Thus, it suffices to approximate p_i to within additive error $\frac{\epsilon}{2n}(1-2\eta)$, which we can do from samples.

This type of estimation to within additive error is formalized by the statistical query model, where we can ask for expectations of bounded functions of the dataset up to additive error. Formally, we may give the SQ oracle a function $h : X \times \{0,1\} \to [0,1]$ and a tolerance $\tau > 0$, and the oracle will respond with $\mathbb{E}_{x \sim D}[h(x)] \pm \tau$. We require h to be polytime computable and $\tau \geq \frac{1}{\text{poly}}$.

The SQ model is in fact extremely general. Many (almost all) known learning algorithms can be implemented with SQ. For example, gradient descent:

$$L(w) = \mathbb{E}_{x,y}[\ell(x,y,w)] \tag{4}$$

$$\nabla_w L(w) = \nabla_w \mathbb{E}_{x,y}[\ell(x,y,w)] \tag{5}$$

$$= \mathbb{E}_{x,y}[\nabla_w(\ell(x,y,w))], \tag{6}$$

and the last equality is an SQ output.

Theorem 1.1

If a concept class is SQ-learnable, then it is PAC learnable with random classification noise.

Proof: We are trying to estimate $\Pr[h(x, f(x) = 1]$. Let $C = \{x : h(x, 0) = h(x, 1)\}$ be the set of clean examples and $N = \{x : h(x, 0) \neq (x, 1)\}$ be the set of noisy examples. Then

$$\Pr[h(x, f(x)) = 1] = \Pr[h(x, f(x)) = 1 \text{ and } x \in C] + \Pr[h(x, f(x)) = 1 \text{ and } x \in N]. \tag{7}$$

We can estimate $\Pr[x \in C]$ since it is not affected by noise. Then by the same argument as the OR example, if $q = \Pr_\eta[\ell(x) = 0 | x \in N]$,

$$\Pr[h(x, f(x)) = 1 | x \in N] = \frac{q - \eta}{1 - 2\eta}. \tag{8}$$

so estimating the left side to within additive error τ will enable us to estimate q to within additive error $\tau(1 - 2\eta)$. ■